Application of Confidence Intervals to Text-Based Social Network Construction

Julie Paynter, Ian McCulloh, John Graham
United States Military Academy
West Point, NY 10996

Abstract

With the increasing importance of gathering intelligence on insurgent and terrorist groups, social network analysis (SNA) has become an important analytic tool. SNA is the mathematical methodology of quantifying connections between individuals and groups. This research is focused on the concept of centrality, which is a mathematical process of determining which node in a network is the most central, or connected. Thematic, or intangible, relationships consist of entities that are not directly connected, but who share similar ideologies. While the concept of centrality – the most connected node – remains the same, the question becomes how to determine if two nodes are connected where a tangible relationship is not present. To determine if there is a connection, t-confidence intervals are constructed for each entity. If they share overlapping confidence intervals, they are connected. The connection is weighted based on a scaled difference between the means of the confidence intervals. The final network consists of nodes connected only across all of the chosen intangible or thematic fields. Analysis of the network is conducted using measures of degree centrality. This paper proposes a new algorithm for determining connection between nodes in a thematic network, using an analysis of radical jihadist writings to demonstrate the applicability of the method.

KEY WORDS: Network, node, centrality, theme, relationship, t-test, geometric mean.

Introduction

A Social Network is a mathematical quantification of connections between groups or people. Social Networks are based on the idea that rational actors are interdependent beings (Wasserman, et. al., 1994). That is, no one person or event can exist in a vacuum. From each person and each event there must invariably be ties or links to other people and events. Those ties can be pictorially depicted as connections (relationships) in a network where the people or events are nodes. Behind each graphically displayed network is a matrix of connections – each node in the chart receives a row and column, and the relationship between two nodes (either

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Application of Confidence Intervals to Text-Based Social Network Construction	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Military Academy,West Point,NY,10996**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

binary or valued) is indicated in the correct intersection in the matrix. From such charts, and the matrices upon which they are developed, network analysts can determine central actors or key events. They use known relationships to gain information or to exploit other members of the network. For example, a massive, hand-drawn and catalogued social network which led U.S. forces to Saddam Hussein.

As we have seen on the battlegrounds of the Global War on Terror (GWOT), the focus of modern war is not kinetic. Instead, commanders at all levels must understand the "human terrain" of their area of responsibility. Much of the information we can use to understand local culture, history, opinion and leadership is in text form: newspaper, memorandum, religious declarations, letters and a multitude of other sources. Reading all of the available information is time consuming. Finding connections and drawing conclusions from the texts is even more challenging. Social networking is a solution, but the current methods do not provide a way to compare key personnel, groups or locations across multiple fields. Analysts are limited to looking at only one area – such as citation analysis, where connections are made when authors cite each other. This paper will explain a method of finding connections across multiple fields.

Literature Review

A new algorithm for constructing a social network across multiple fields will be demonstrated on an example data set of radical Islamist writers. The data set is composed of approximately 250 translated texts provided by the Combating Terrorism Center at West Point (CTC). From the 250 texts, 92 were chosen that were published by a selected set of fifteen authors. The authors were chosen based on two criteria: having more than two texts in the original data set, and not being well-known. This means that Bin Laden and Zarqawi, among

others, were not analyzed because conclusions reached about them were not likely to be new or useful to the CTC. The fifteen authors are (in alphabetical order by first letter of first word): Abd-al-Aziz al-Muqrin, Abu-Maysarah al-Iraqi, Muhammad Alshareef, Muhammad Naasirud Deen al , Shaykh Abdul Qadir Bin Abdul Aziz, Shaykh 'Abdul-'Azeez Ibn Baaz, Shaykh Abdullah Azzam, Shaykh Abu Basir At-Tartusi, Shaykh Abu Mohammed al-Maqdisi, Shaykh Hammoud bin Uqlaa' Ash-Shuaibi, Shaykh Naasir ibn Hamad al-Fahd, Shaykh Rabee' ibn Haadee al-Madkhalee, Sheik Muhammad Ibrahim Al-Madhi, Sheikh Salman al-Awdah, and Sheikh Yussef Al-Qaradhawi. The texts were obtained mainly from the Foreign Broadcast Information Service and the Middle East Media Research Institute, with a few obtained from various websites. After cataloging all of the administrative data about each text and cleansing the texts to account for varied spellings and proper nouns, they were analyzed using Crawdad, a text analysis software developed at Arizona State University to perform centering resonance analysis (Corman, et. al., 2005). The software assigns an influence value to each word in the text. The CTC developed themes to look for in the texts; each theme is composed of a list of words that relate to that theme. The eight themes are shown below in Table 1.

TABLE 1 ABOUT HERE

Each text was assigned a score for each theme, where the score consisted of the sum of the influence values of each word in the theme. Once this process was complete, the data was compiled so that each author had a list of theme scores: one for each theme from each text. Using basic descriptive statistics, the following theme rankings were produced, these rankings will be compared to the results obtained using the themed network algorithm later in the paper. Table 2

shows theme ranks for the fifteen authors in each of the eight themes, as well as an average theme rank.

TABLE 2 ABOUT HERE

A theme rank of 1 indicates that the author had the highest mean theme score for the given theme, while a score of 15 indicates the lowest mean score. The "Average Rank" column is simply an average of the eight theme scores for each author, while the "Overall" column denotes the author's average theme rank. Al-Fahd has the highest average rank (assuming that 1 is the highest rank), thus he is the overall number 1 author.

Further analysis of the data set was completed using a plagiarism check. This software program checks texts for matching word strings of five words or more. Any two authors with four or more connections were drawn as connected nodes in a network. The network is shown in Figure 1.

FIGURE 1 ABOUT HERE

This method has inherent flaws. First, it does not account for context of the words in the text – it is possible to say the same thing as another person and to mean something entirely different. Second, it oversimplifies by assigning word matches to rote statements such as the Koranic intonations that begin many of the texts. Finally, and most importantly, the plagiarism check is weighted very heavily towards prolific authors. Al-Awdah is the most central person in this network, meaning that he has the most connections, and he is also the author with the most

writing (in terms of length, not number of texts). Maqdisi is the next most central, and he wrote the second highest amount. In fact, there is almost a one-to-one correspondence between the centrality rank of each author and their rank in terms of how much they wrote. Clearly this method is flawed in terms of providing meaningful results. The algorithm proposed in this paper improves upon these current methods of constructing a social network.

Methodology

Various mathematical procedures have been used to find fissures in the Islamic extremist ideological movement based on jihadist texts. Originally using ANOVA, and then Kruskal-Wallis, it was found that neither method was capable of producing meaningful analysis. This newly proposed algorithm was developed as an alternative method for finding connections across multiple categories using mathematically sound procedures. The algorithm uses t-confidence intervals and geometric means to construct a network across multiple categories, or fields. Starting with a set of scores for each node in the future network, we construct confidence intervals, determine interrelation scores based on the intervals, and construct a matrix. The resultant network is a representation of the geometric means of each pair of node's relationships in each field.

Given a set of nodes to analyze – radical Islamic authors are one example - one must create a set of scores for each node. Although there are various ways to do this, the example data set used in this paper obtained theme scores from Crawdad, a text analysis software program, for each author and each text. With a set of scores for each node, it is possible to construct confidence intervals that approximate the nodes normal range of values in the specified field.

The primary drawback to confidence intervals is their symmetry. This means that a grossly asymmetric data set will not actually meet the stated level of confidence. For example, given a data set with an arbitrary left end-point, such as hourly wages in the U.S., the data will be skewed to the right. Skewness may cause the confidence level to be closer to 85% or 90% than to 95%, but if the data set has greater than thirty points, this becomes less of a problem because of the central limit theorem. Thus, a 95% confidence interval should produce an acceptable range in which that node's mean score for a certain field is likely to fall. For the example set in this paper, there may be some bias because there were between two and eighteen texts per author.

Using the set of scores for each node, a confidence interval is calculated for each node in each field. The confidence interval is given by,

$$\bar{x} \pm t_{.025,n-1} \cdot \frac{s}{\sqrt{n}}, \tag{1}$$

where $\bar{x}$ is the sample mean of the theme score for a particular author, $t_{.025,n-1}$ is the t-statistic for a 95% confidence interval with $n-1$ degrees of freedom, s is the sample standard deviation of an author's text scores, and n is the number of texts by an author.

Once the confidence intervals are constructed, they are used to determine if two nodes are similar in each field. If two nodes do not have overlapping confidence intervals, then they receive an interrelation score of 0. If two nodes have overlapping confidence intervals in a field, then they receive an interrelation score given by,

$$s_{i,j} = \frac{MD - AD}{MD}, \tag{2}$$

where MD is the maximum possible difference between any two means within the field, AD is the actual difference between the two nodes means, and $s_{i,j}$ denotes the matrix entry for the two

nodes (*i* and *j*). This relationship score is scaled so that a maximum difference of means will produce zero, while a difference of zero produces a score of one. Scaling is necessary to ensure that data across all fields can be compared without bias.

Each relationship score is placed in a matrix for each field. Table 3 shows an example matrix.

TABLE 3 ABOUT HERE

Once matrices are constructed for each field, the newly proposed algorithm provides a method to combine them. The geometric mean of each pair of nodes for each field is taken, and the resultant number is placed in a new square matrix, again with, the nodes as the rows and columns. The formula for the geometric mean is:

$$\left(\prod_{l=1}^{f} a_l\right)^{\frac{1}{f}} \tag{3}$$

where f is the number of fields and a_l is the relationship score in field l for the two nodes in question. If any pair of nodes has an interrelation score of zero for any field, the geometric mean ensures that the combined score is multiplied by zero, thus forcing their overall relationship score to zero. The reason behind this method is that if two nodes are connected in all themes, they are much more similar than nodes that are only connected in a few fields. This does leave open the possibility that two nodes may be connected in all but one field and register an overall score of zero. However, different risk functions could be investigated for other applications of themed network analysis.

The final result is a square matrix containing the overall relationship scores of each pair of nodes across all measured fields. From this matrix, computer software is used to construct the graphical network.

Results

The network constructed using the newly proposed algorithm differed from those produced using either the plagiarism check or the average theme ranks. The final square matrix from the new method is displayed in Table 4.

TABLE 4 ABOUT HERE

The matrix produces a more insightful network and is shown in Figure 2.

FIGURE 2 ABOUT HERE

The newly proposed algorithm, like the average theme ranks, places Al-Fahd as the most central figure. However, the average theme ranks approach does not account for connections between the authors. This makes the measure excellent at determining which themes an author focuses on, but not at determining which author is the most influential within the group. A significant advantage of the newly proposed algorithm is that it does tell us who is the most influential: Al-Fahd, with Maqdisi second. The numerical representation of this (based on weighted degree centrality) is shown in Table 5.

TABLE 5 ABOUT HERE

It can be seen in Table 5 that Al Fahd is the most influential author in the group. Al-Fahd is a radical Saudi cleric who was very influential in the jihad movement in the 1990s (Brachman, 2006), which explains his influence in the network. Maqdisi, the second most influential, was Zarqawi's mentor when the two were imprisoned together (Al-Zaydi, 2005). Clearly, both men have been and still are influential ideologues within the jihadist movement. The largest confirmation of the method, however, is that Ibn Baaz is the third most influential figure – he was not in the top five using either average theme ranks or the plagiarism check. Ibn Baaz, also known as Bin Baz, has been described as one of the "founding fathers" of the jihadist movement; though not an advocate of violence or offensive jihad, his writings laid the theological groundwork for the current Islamic Fundamentalist jihadists (Brachman, 2006). In any list of influential Islamic writers, Ibn Baaz would be included, and his inclusion here is reassurance that the confidence interval method is valid. Also corroborating the proposed algorithm is the placement of Madkhalee and Al-Albanee, two moderates who advocate peaceful means of advancing their fundamentalist ideology, as outliers. Within the selected group, they are at odds with the majority, once again evidencing that the algorithm produced results consistent with reality.

Conclusion

Methods of constructing author theme social networks that simply rank the theme scores of each author ignore the possibility and implication of connections between the authors. A simple word matching program cannot account for content and is greatly affected by the amount

of data to analyze. Themed analysis, however, using confidence intervals to discern similarity, can provide an improved measure of connectivity. Furthermore, using confidence intervals negates the effect of voluminous or scanty writings. By analyzing an example data set of jihadist authors, the newly proposed algorithm demonstrates its ability to provide analysis of either abstract or concrete data, to find linkages across several different fields, and to find intangible connections such as the thematic relationships previously discussed. In short, this algorithm provides an improved method for finding connections in large amounts of textual data.

References

Al-Zaydi, Mshari. "Al Muqrin: Al-Qaeda's Angel of Death: From goalkeeper to Al-Qaeda"s Death keeper; the story of the making of a terrorist." *asharq alawsat*, 20 June 2005. Available from http://www.asharqalawsat.com/english/news.asp?section=3&id=511. Accessed July, 2005.

Brachman, Jarret, Senior Research Fellow, The United States Military Academy Combating Terrorism Center, West Point, New York. March, 2006.

Corman, Steven and Dooley, K. Hugh Downs School of Human Communication, Arizona State University. August, 2005.

Wasserman, Stanley and Katherine Faust. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press, 1994.

THEMES							
ISLAM	JIHAD	SALAF	INFIDEL	FOREIGNERS	SHEIKH	BATTLEGROUNDS	JEWS
allah	al_jihad	salaf	infidel	united_states	shaykh	Afghanistan	jews
religion	mujahid	sunnah	apostate	government		bosnia	zionists
islam	attack	sallam	heretic	al-Saud		two-rivers	usury
muslim	raid		kuffr	Australia		iraq	israel
ummah	defense		taghoot	Britain		palestine	
brother	plane		idol	Spain			
book	bombing			Italy			
messenger	operation			France			
prophet	clash						
mohammad	fight						
	conflict						

Table 1 – Eight Themes For Analysis

Theme Ranks										
Author	Islam	jihad	salaf	infidel	foreigners	battlegrounds	sheikh	jew	Average Rank	Overall
al-Fahd	10	5	6	3	7	6	2	9	5.57	1
Mugrin	6	4	12	6	1	4	11	9	6.29	2
Shuaibi	11	9	11	1	6	5	3	9	6.57	3
Azzam	12	2	10	7	8	3	7	8	7.00	4
Maqdisi	9	1	8	8	5	9	10	4	7.14	5
al-Iraqi	8	6	13	5	9	1	11	9	7.57	6
At-Tartusi	14	8	4	2	15	10	1	5	7.71	7
Abdul Aziz	5	10	9	4	10	12	5	6	7.86	8
Madhi	2	7	13	12	3	7	11	2	7.86	9
Qaradhawi	15	3	13	9	11	2	4	1	8.14	10
Alshareef	3	14	3	14	2	11	11	3	8.29	11
Madkhalee	4	13	2	11	12	12	6	9	8.57	12
Al-Awdah	13	11	7	10	4	8	8	7	8.71	13
al Albanee	1	14	1	14	14	12	11	9	9.57	14
Ibn Baaz	7	12	5	13	13	12	9	9	10.14	15

Table 2 – Theme Ranks of Fifteen Jihadist Authors

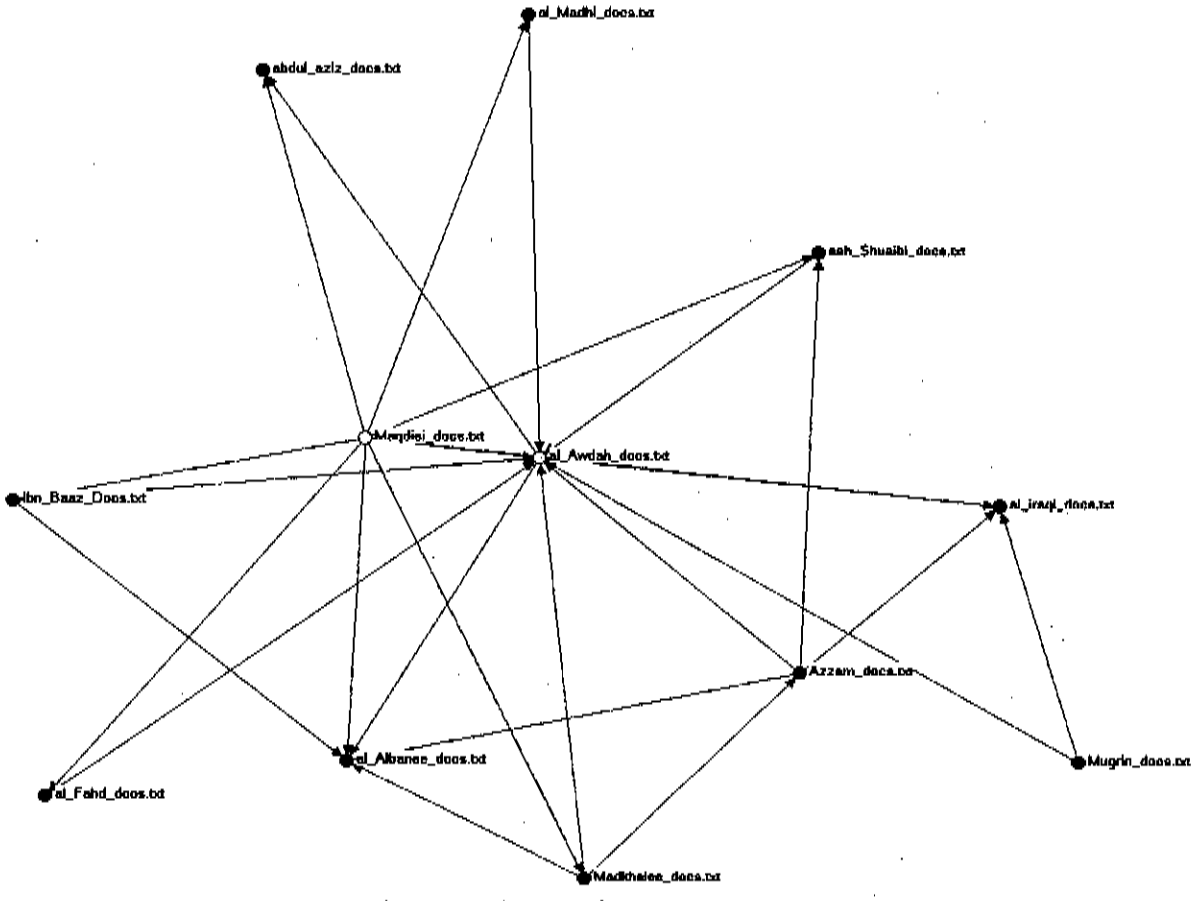

Figure 1 – Plagiarism Check Network Diagram

Field			
	Node A	Node B	Node C
Node A	A,A	A,B	A,C
Node B	B,A	B,B	B,C
Node C	C,A	C,B	C,C

Table 3 – Example Field Matrix of Interrelation Scores

Overall Theme Scores															
	Mugrin	al-Iraqi	Alshareef	al Albanee	Ibn Baaz	Abdul Aziz	Azzam	At Tartusi	Maqdisi	Shualbi	Al-Fahd	Madkhalee	Madhi	Al-Awdah	Qaradhawi
Mugrin	1.00000	0.76695	0.00000	0.00000	0.00000	0.00000	0.84938	0.00000	0.84852	0.80676	0.83939	0.00000	0.84403	0.00000	0.00000
al-Iraqi	0.76695	1.00000	0.51748	0.00000	0.00000	0.00000	0.84449	0.00000	0.69722	0.82516	0.81203	0.00000	0.72532	0.00000	0.00000
Alshareef	0.00000	0.51748	1.00000	0.75690	0.83688	0.00000	0.00000	0.00000	0.00000	0.00000	0.77599	0.00000	0.94816	0.00000	0.00000
al Albanee	0.00000	0.00000	0.75690	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.90076	0.00000	0.00000	0.00000
Ibn Baaz	0.00000	0.00000	0.83688	0.00000	1.00000	0.91174	0.82297	0.78024	0.80594	0.90168	0.91619	0.00000	0.86383	0.87589	0.89418
Abdul Aziz	0.00000	0.00000	0.00000	0.00000	0.91174	1.00000	0.00000	0.00000	0.73681	0.82157	0.85487	0.95733	0.88681	0.94896	0.00000
Azzam	0.84938	0.84449	0.00000	0.00000	0.82297	0.00000	1.00000	0.59977	0.93159	0.81634	0.89227	0.00000	0.79010	0.00000	0.83695
At Tartusi	0.00000	0.00000	0.00000	0.00000	0.78024	0.00000	0.59977	1.00000	0.52446	0.81676	0.82689	0.00000	0.00000	0.00000	0.00000
Maqdisi	0.84852	0.69722	0.00000	0.00000	0.80594	0.73681	0.93159	0.52446	1.00000	0.77203	0.86424	0.00000	0.82544	0.76406	0.77915
Shualbi	0.80676	0.82516	0.00000	0.00000	0.90168	0.82157	0.81634	0.81676	0.77203	1.00000	0.92896	0.00000	0.87030	0.84583	0.00000
Al-Fahd	0.83939	0.81203	0.77599	0.00000	0.91619	0.85487	0.89227	0.82689	0.86424	0.92896	1.00000	0.00000	0.86821	0.86983	0.00000
Madkhalee	0.00000	0.00000	0.00000	0.90076	0.00000	0.95733	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
Madhi	0.84403	0.72532	0.94816	0.00000	0.86383	0.88681	0.79010	0.00000	0.82544	0.87030	0.86821	0.00000	1.00000	0.00000	0.00000
Al-Awdah	0.00000	0.00000	0.00000	0.00000	0.87589	0.94896	0.00000	0.00000	0.76406	0.84583	0.86983	0.00000	0.00000	1.00000	0.00000
Qaradhawi	0.00000	0.00000	0.00000	0.00000	0.89418	0.00000	0.83695	0.00000	0.77915	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000

Table 4 – Overall Theme Interrelation Scores from Proposed Algorithm

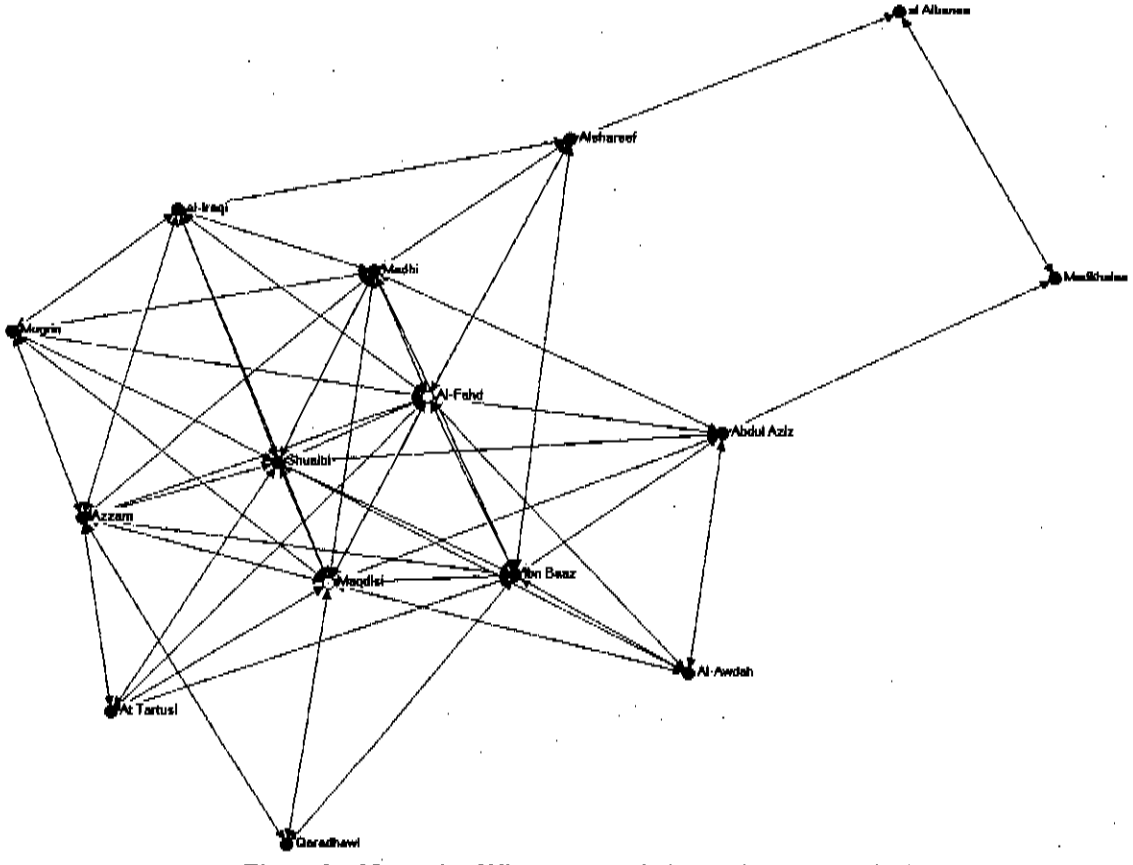

Figure 2 – Network of Theme Interrelations using Proposed Algorithm

Weighted			
	Texts	Degree	NrmDegree
Al-Fahd	8	9.389	70.053
Maqdisi	6	8.549	63.789
Ibn Baaz	10	8.41	62.745
Shuaibi	3	7.606	56.753
Madhi	4	7.26	54.17
Azzam	4	7.185	53.608
Abdul Aziz	4	5.818	43.41
al-Iraqi	7	5.189	38.714
Mugrin	10	4.955	36.971
Al-Awdah	16	4.105	30.625
Alshareef	2	3.833	28.602
At Tartusi	2	3.55	26.489
Qaradhawi	7	2.112	15.76
Madkhalee	7	1.858	13.864
al Albanee	2	1.658	12.368

Table 5 – Weighted Degree Centrality of Theme Interrelation Network